REMARKS

Applicant has carefully reviewed and considered the Examiner's Action mailed July 13, 2007. Reconsideration is respectfully requested in view of the foregoing amendments and the comments set forth below.

By this Amendment, claims 1-2, 4, 8, 11, 14, 16-18, and 20 are amended to define the invention more precisely. The amendments are thought to overcome the grounds for rejection for the reasons given below. Support for the amendments can be found in the specification as noted below. Accordingly, claims 1-20 are pending in the present application.

Claim 20 was rejected under 35 U.S.C. §101 for the reasons set forth in paragraph 4 of the Action. Applicant has amended claim 20 to recite a "tangible" machine-readable medium. Thus, claim 20 cannot be interpreted "as merely a magnetic carrier wave", as suggested in the Action. Accordingly, it is believed that claim 20 is directed to statutory subject matter and withdrawal of the rejection under 35 U.S.C. §101 is respectfully requested.

Claim 1, 13 and 20 were rejected under 35 U.S.C. §102(b) as being anticipated by Nagata (A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A N Best Search Algorithm). This rejection is traversed.

Claims 1 and 13 of the present application are directed to a morphological analyzer and a method of morphological analysis. A part-of-speech n-gram model is a linguistic model that breaks a sentence into strings of n consecutive morphemes, assigns a part of speech to each morpheme, and uses this information to calculate the probability that a given morpheme will appear in a given context. The claimed invention attempts to

resolve ambiguity in the morphological analysis of sentences by using different types of part-of-speech n-gram models and combining the probabilities they produce. The different types of models referred to in the specification and claims of the present application are a hierarchical part-of-speech n-gram model, a lexicalized part-of-speech n-gram model, and a class part-of-speech n-gram model.

A simple part-of-speech N-gram model considers the probability of occurrence of a morpheme (word) w_i as the i-th morpheme in a sentence if the i-th morpheme is assumed to have part of speech t_i, and the probability of occurrence of this part of speech t_i following a string of unspecified words with parts of speech $t_{i-N+1},\ldots, t_{i-1}$, as shown in equation (3) on page 10 of the originally filed specification of the present application. This model is based on the combination of a morpheme and its part of speech, and the combination of a string of consecutive parts of speech.

In the hierarchical part-of-speech n-gram model, a morpheme is tagged with its part of speech, and for some parts of speech, the tag also indicates a particular form of the part of speech (an English language example of the these forms might be the different tense and plural/singular form of a verb). In the probability calculations, parts of speech and their forms are considered separately, as shown by equation (7) on page 10 of the originally filed specification of the present application. This model is based on the combination of a morpheme and its part of speech, the combination of a part of speech and its particular form, and the combination of a string of consecutive parts of speech.

In a lexicalized part-of-speech model, occurrences of specific words with specific parts of speech are considered, as in equations (4) to (6) on page 10 of the originally filed specification of the present application, which describe three different lexicalized part-of-

speech models. These models are based on combinations of morphemes and their parts of speech, and in one model on the combination of a morpheme, its part of speech, and the preceding string of parts of speech.

A class part-of-speech n-gram model groups morphemes into clusters or classes, and uses combinations of a particular class of morpheme with a particular part of speech, as in equations (21) and (22) on page 20 of the originally filed specification of the present application, which describe two different class part-of-speech models. These models are based on the combination of a morpheme and its part of speech and the combination of a class and its part of speech.

Claim 1 was rejected over the Nagata reference, which uses a simple tri-POS model (a part-of-speech 3-gram model) including information about the different forms that a part of speech may take, as described in the first footnote on page 202. The Nagata reference discloses an interpolated estimation process that interpolates trigram, bigram, unigram, and zerogram relative frequencies (i.e., it uses part-of-speech n-gram models with n = 3, 2, 1, and 0). However, claims 1 and 13 have been amended to further define that "at least two of the part-of-speech n-gram models [are] based on mutually different types of morphological information". That is, it is clear that the claimed different types of part-of-speech n-gram models must include at least two models based on different types of morphological information (structure and form of linguistics), not just different values of n as disclosed by Nagata. Consequently, Nagata cannot anticipate claims 1, 13 and 20 because it fails to disclose each and every feature of the claimed invention. Withdrawal of the rejection under 35 U.S.C. §102(b) is respectfully requested.

Claims 2-3 and 14-15 were rejected under 35 U.S.C. §103(a) as being unpatentable over Nagata. This rejection is respectfully traversed.

Amended claims 2 and 14 state that the part-of-speech n-gram model including information about forms of part of speech is a hierarchical model that treats parts of speech and their forms at different hierarchical levels, as is evident from the separation of t_i^{form} from t_i^{pos} in equation (7) on page 10 of the originally filed specification of the present application. Claims 3 and 15 (Original) gives a verbal statement of equation (7). It was the Examiner's position that claims 2-3 and 14-15 were obvious from column 1, line 6 on page 202 in the Nagata reference. Claims 2 and 14 have been amended to specify that parts of speech and their forms [of a hierarchical part-of-speech n-gram model] must be on different hierarchical levels. In footnote 1 on page 202, Nagata explains that he uses 120 part-of-speech tags, each tag listing a part of speech, its conjugation type, and its conjugation form. The equations in column 1 on page 202 of Nagata make it clear that these tags t_i are treated as single units; there is no separation of a part of speech (t_i^{pos}) from its form (t_i^{form}) as in the invention set forth in claims 2 and 14 of the present application.

Thus, it follows that Nagata's model treats parts of speech and theirs forms on the same hierarchical level, failing to meet the limitations of amended claims 2 and 14. It is submitted that one of ordinary skill in the art reading Nagata would not have considered separating parts of speech and their forms as the morphological task disclosed by Nagata is concerned with "finding a set of word segmentation and parts of speech assignment that maximize the joint probability of word sequence and tag sequence P(W,T)." See page 201, column 2, lines 30-33 of Nagata.

Although Nagata does use a type of hierarchical scheme, it is the trigram-bigram-unigram-zerogram hierarchy defined by equation (5) of Nagata: an n-gram hierarchy based on the value of n. There is no suggestion in Nagata of treating parts of speech and their forms on different hierarchical levels, as required by amended claims 2 and 14 of the present application.

Equations (3) and (4) of Nagata (cited by the Examiner) fail to match the mathematical expressions defined in claims 3 and 15 of the present application. For example, the expression $P(t_i^{form}|t_i^{pos})$, defined in claims 3 and 15 of the present application as the conditional probability of occurrence of part of speech t_i^{pos} in form t_i^{form}, does not match the expression $P(w_i|t_i)$, which Nagata defines in the last sentence on page 201 as the probability of output of a word (w_i) if it is assumed that the word has a given part of speech (t_i). Accordingly, claims 2-3 and 14-15 are patentable over Nagata at least for the reasons given above. Withdrawal of the rejection under 35 U.S.C. §103(a) is respectfully requested.

Claims 4-7 and 16 were rejected under 35 U.S.C. §103(a) as being unpatentable over Nagata in view of Pla et al. (Improving Part of Speech Tagging using Lexicalized HMMs - hereinafter referred to as "Pla"). This rejection is traversed.

Amended claims 4 and 16 specify use of a lexicalized part-of-speech n-gram model that lexicalizes all words. Claims 4-7 specify three specific models, corresponding to equations (4)-(6) in the specification of the present application.

Nagata teaches the use of a part-of-speech n-gram model. Pla teaches the use of a lexicalized hidden Markov model (HMM) for improving precision of part-of-speech tagging. The Examiner combines these teachings to obtain a lexicalized part-of-speech

n-gram model. However, Pla does not disclose the feature argued above that is missing from Nagata: "at least two of the part-of-speech n-gram models [are] based on mutually different types of morphological information".

Pla describes his lexicalization scheme on page 6. A hidden Markov model specifies the probabilities that various words will be emitted in various states. Pla's lexicalization scheme identifies certain specialized words and the states from which they are emitted, and splits each of these states into two states, one of which emits the specialized word, the other of which emits all other emitted words. In the example spanning pages 7 and 8 of Pla, the 'IN' state that emits a subordinating conjunction or preposition is split into a 'that IN' state that always emits the word 'that' and a general 'IN' state that emits any other subordinating conjunction or preposition. Accordingly, Pla's model is lexicalized only with respect to certain specialized words. In contrast, the lexicalized part-of-speech n-gram models of the claimed invention are lexicalized with respect to all words. Equations (4)-(6) on page 10 of the present invention and represented in words by claims 5-7 of the present application do not single out any specialized words or give them special treatment.

The mathematical description of Pla cited by the Examiner, starting in line 23 on page 7 of Pla, discloses lexical probabilities $P(w_i \mid c_i)$ that is similar to conditional probability shown on the left side of equation (10) on page 13 of the present specification. Pla further discloses the lexical probabilities are obtained by dividing the frequency of the pair $\langle w_i, c_i \rangle$ by the frequency of the category c_i that is similar to the right side of equation (12) on page 13 of the present specification. The equations of Pla describe how certain probabilities are calculated from a corpus of text. They do not,

however, describe how the calculated probabilities are combined with other probabilities to create a lexicalized part-of-speech n-gram model of the type as recited in any of equations (4)-(6) and defined in any of claims 5-7 of the present application. Furthermore, attaching an equation similar to the left side of equation (10) of the present application to an equation similar to the right side of equation (12) of the present application will not produce the same result as either equation (10) or equation (12) in the present specification.

Claims 8 and 17 were rejected under 35 U.S.C. §103(a) as being unpatentable over Nagata in view of U.S. Patent Application Publication No. 2003/0046078 to Abrego et al. (hereinafter referred to as "Abrego"). Claims 9-11 and 18 were rejected under 35 U.S.C. §103(a) as being unpatentable over Nagata and Abrego and further in view of Pla. These rejections are traversed.

Amended claim 8 and depending claims 9-11 recite the use of a class part-of-speech n-gram model employing classes obtained by clustering and describe specific models and clustering methods. The Examiner rejected claim 8 over Nagata and Abrego, using Abrego to supply the class model lacking in Nagata.

However, in paragraphs 0022-0023 of Abrego, it is disclosed that Abrego has text formatted for reading convenience from the actual output from a morpho-syntactic analyzer and a conventional expert system to create word classes. There is no disclosure of obtaining the classes by clustering, as required by claims 8-11 and 17 of the present application. The classes described by Abrego, such as 'noun--building' are refinements of conventional parts of speech on the basis of expert knowledge, such as knowledge that the nouns 'house', 'office', and 'church' are types of buildings. Paragraph 0038 of

Abrego cited by the Examiner does not disclose a "class part-of-speech n-gram model employing classes obtained by clustering", as required by claims 8 and 17 of the present application.

Amended claims 11 and 18 of the present application recites training the class part-of-speech n-gram model from both a part-of-speech tagged corpus and a part-of-speech untagged corpus, using clustering parameters obtained from the part-of-speech untagged corpus to cluster morphemes in the part-of-speech tagged corpus. This is described in the specification from line 8 on page 21 to line 7 on page 22. Note that lines 20-21 on page 21 of the present specification state that words are assigned to classes by using only the word information in the corpus, in contrast to Abrego's use of expert knowledge.

The Examiner rejected claims 11 and 18 on the basis of Nagata, Abrego, and Pla, citing Pla's statement that "the learning process of the parameters in equation 2 can be learned from labeled corpora ... or from unlabeled corpus" (page 7, line 15 of Pla), but the parameters in equation 2 (given by Pla on page 5) are the conditional probabilities for the occurrence of words w_i in states c_i to which part-of-speech tags are associated (page 5, lines 12-18 of Pla). Thus, the parameters disclosed by Pla are not classes of words.

Pla states that when a labeled corpus is used, the model (in this case, a hidden Markov model) is trained from the observed frequencies, and that when an unlabeled corpus is used, the model is trained by the Baum-Welch algorithm (page 7, lines 15-22). This does not suggest either the combined use of both a tagged corpus and an untagged corpus, or the use of an untagged corpus to obtain clustering parameters, as recited in amended claims 11 and 18.

By the same token, amended claim 11 does not advocate the use of use of an untagged corpus to calculate the probabilities in a model, which is what Pla suggests. Instead, claims 11 and 18 recite that the class part-of-speech n-gram model is trained from both a part-of-speech tagged corpus and a part-of-speech untagged corpus where clustering parameters obtained from the part-of-speech untagged corpus are used to cluster morphemes in the part-of-speech tagged corpus. It is respectfully submitted that the claimed use of class models reduces the occurrence of unwanted side-effects that class models tend to produce, as noted in the sentence bridging pages 20 and 21 in the specification of the present application. This feature is not rendered obvious by any combination of the prior art of record.

Claim 12 was rejected under 35 U.S.C. §103(a) as being unpatentable over Nagata in view of Siu (Variable N-Grams and Extensions for Conversational Speech Modeling). This rejection is respectfully traversed.

Claim 12 recites the weighting of the part-of-speech n-gram models by weights calculated by a leave-one-out method. The originally-filed specification describes such a method in equation (18) on page 16 and illustrates the same in the flowcharts in Figs. 4 and 11 of the present application.

The Examiner rejected claim 12 over the combination of Nagata and Siu, using Siu to supply the leave-one-out method. However Siu does not disclose "at least two of the part-of-speech n-gram models [are] based on mutually different types of morphological information", which is missing from Nagata and recited in independent claim 1. Further, the leave-one-out likelihood disclosed by Siu estimates the distributions used in the distance measures as explained in column 2 of page 69 of that

article. That is, Siu uses the LOO likelihood to estimate probability distributions, specifically, the probability $p_i(x)$ that a word x will occur at a node n_i in an n-gram model. There is no suggestion of using a leave-one-out method to calculate weights for use in combining different n-gram models. Nor is there any reason to believe that a method used by Siu to evaluate the effect of node pruning and merging within one n-gram model (a process of modifying the model) would be of any use in combining different n-gram models (not a modification process).

Consequently, even if Siu and Nagata were to be combined as the Examiner proposes, the result would be to use the leave-one-out likelihood to calculate the weights q_3, q_2, q_1, q_0 in Nagata's equation (5), which are used to combine trigram, bigram, unigram, and zerogram models, and these models are not based on different types of morphological information as required by amended claim 1, on which claim 12 depends.

Claim 19 appears to have been rejected using the combination of Nagata and Siu on page 17 of the Action. Claim 19 recites a similar feature as that of claim 12. Accordingly, claim 19 is not rendered obvious by any combination of Nagata and Siu for the reasons expressed above (independent claim 13 recites similar language as claim 1).

For the above stated reasons, it is submitted that all of the claims are allowable over the prior art of record and are in condition for allowance. Therefore, it is respectfully requested that this application be passed to issuance with claims 1-20 being allowed over the prior art of record.

Should the Examiner believe that a conference would advance the prosecution of this application, he is encouraged to telephone the undersigned counsel to arrange such a conference.

Respectfully submitted,

Date: October 31, 2007

Catherine M. Voorhees
Registration No. 33,074
VENABLE LLP
P.O. Box 34385
Washington, D.C. 20043-9998
Telephone: (202) 344-4000
Telefax: (202) 344-8300

CMV/elw

::ODMA\PCDOCS\DC2DOCS1\903464\1